ina
MÉMOIRE AUGMENTÉE

L'*Intelligence Artificielle* au service de nouvelles opportunités pour *les institutions patrimoniales*

Futurs fantastiques – BnF – 10 décembre 2021

Gautier Poupeau
gpoupeau@ina.fr
@lespetitescases

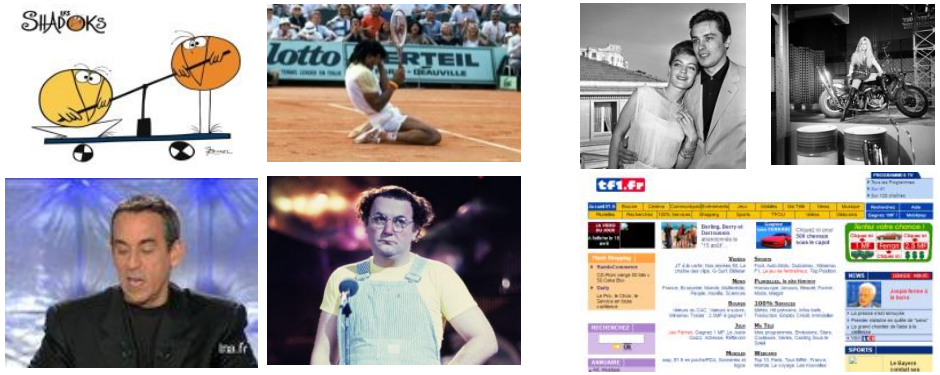
ina
MÉMOIRE
AUGMENTÉE

01.

→ Quelques constats

L'**Institut National de l'Audiovisuel**, c'est…

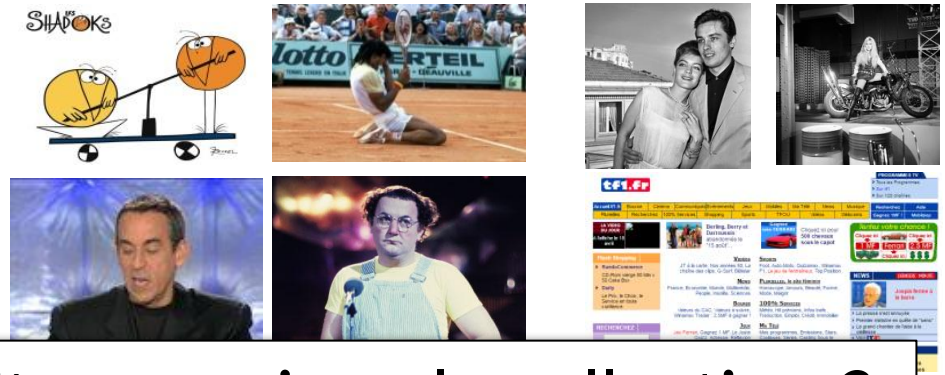
100 km linéaires d'archives conservées sur supports originaux
30 Po d'archives numérisées ou nativement numériques



22 millions d'heures de programmes télé et radio
2 millions de photos et **3.29 Po** d'archives du Web

184 chaînes de télé/radio captées 24h/24 7j/7
Plus de **40 millions** de programmes/sujets télé/radio décrits
pour **59 millions** de diffusions référencées

L'**Institut National de l'Audiovisuel**, c'est…



Au regard des volumétries, comment exploiter au mieux la collection ?

30 Po d'archives numérisées ou nativement numériques

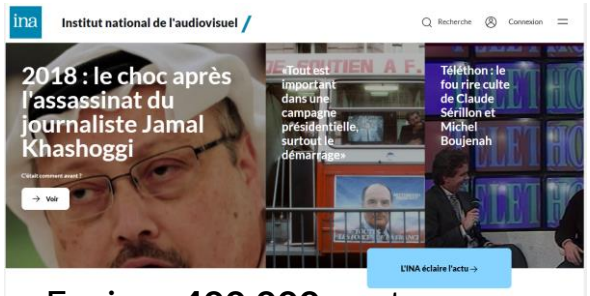
2 millions de photos et **3.29 Po** d'archives du Web

184 chaînes de télé/radio captées 24h/24 7j/7
Plus de **40 millions** de programmes/sujets télé/radio décrits
pour **59 millions** de diffusions référencées

L'**INA**, un *média patrimonial* pour…

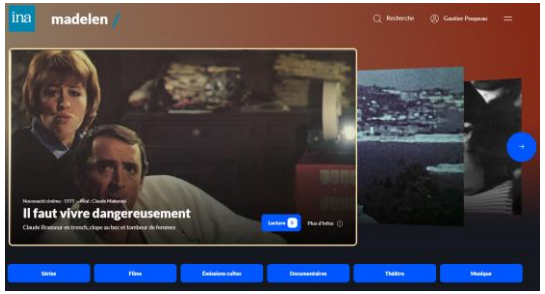
Grand public/B2C

http://www.ina.fr



Environ **400 000** contenus pour **80 000** heures de contenu

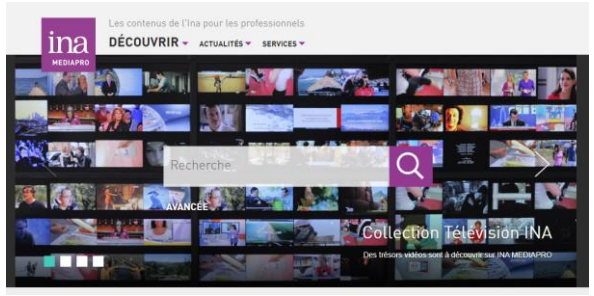
http://madelen.ina.fr



Catalogue SVOD d'environ **13 000** programmes

Professionnels/B2B

https://www.inamediapro.com



Près de **2 millions** d'heures disponibles

https://mediaclip.ina.fr

Environ **500 000** extraits prêts à l'emploi

Etudiants/chercheurs

http://www.inatheque.fr

7 centres donnent accès à tout le fonds
100 centres donnent accès en autonomie à une partie du fonds

L'**INA**, un *média patrimonial* pour…

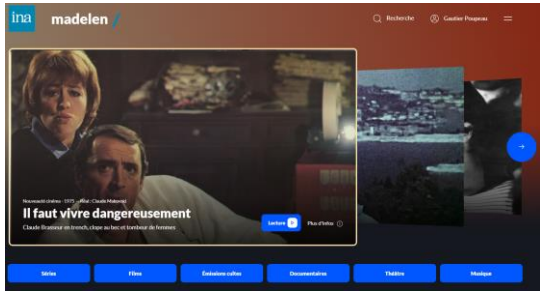
Grand public/B2C

http://www.ina.fr

En

80 000 heures de contenu

http://madelen.ina.fr



Catalogue SVOD d'environ **13 000** programmes

Professionnels/B2B

https://www.inamediapro.com

Près de **2 millions** d'heures disponibles

https://mediaclip.ina.fr

Environ **500 000** extraits prêts à l'emploi

Etudiants/chercheurs

http://www.inatheque.fr

7 centres donnent accès à tout le fonds
100 centres donnent accès en autonomie à une partie du fonds

Mais, malgré nos efforts, ce n'est qu'une goutte d'eau dans l'océan…

Dans le même temps, la *société interrogent les médias* :

fake news, diversité, pluralisme des sujets, présence à l'écran…

Cela doit nous amener à nous interroger….

Comment exploiter et donner du sens à cette *masse de données*,
au-delà de nos missions traditionnelles de *conservation* et d'*accès* aux personnes accréditées ?

Comment *rendre tangibles* ces éléments tout en respectant nos impératifs juridiques ?

En quoi cette masse peut répondre aux *interrogations politiques et sociales* ?

Jusqu'à quel point pouvons-nous rendre cette masse *intelligible et comment* ?

Quel est *notre rôle* dans cette perspective ?

ina
MÉMOIRE AUGMENTÉE

02.

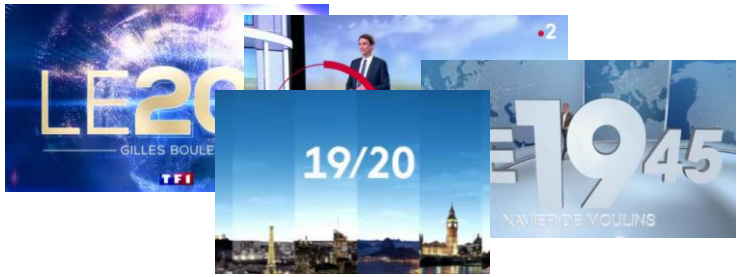
→ Notre proposition

Une rencontre entre…

Des corpus

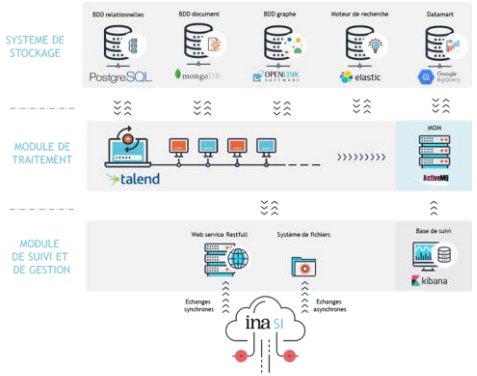
issus de la captation de 184 chaînes de télé et radio pour le dépôt légal

Chaînes d'information continue



Journaux télévisés

Des technologies



Une plateforme de stockage et traitement des données en masse

Des systèmes d'analyse automatique de l'image et du son

Des questions

Sur la diversité dans les médias

Comment se répartit la parole des hommes et des femmes dans les médias ?

Sur l'actualité vue par les médias

Quels sont les personnes, les lieux et les thèmes traités, cités, vus dans les médias ?

Sur le contenu des médias

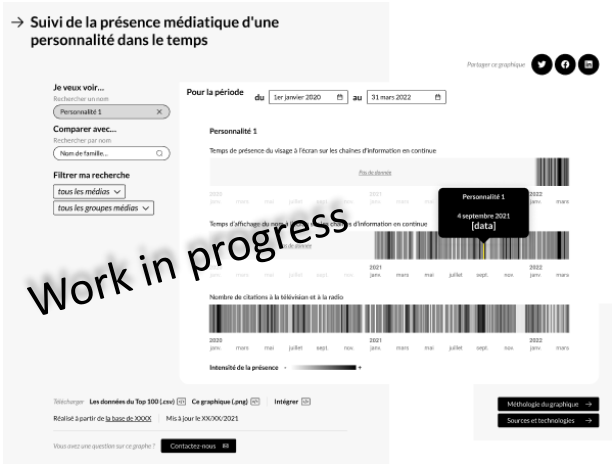
Quel est l'offre de programmes dans les médias ? Comment évolue-t-elle ?

… pour proposer un site d'analyse des médias accessible à tous

Le site sera disponible au cours du 2ème trimestre 2022 à l'adresse https://data.ina.fr

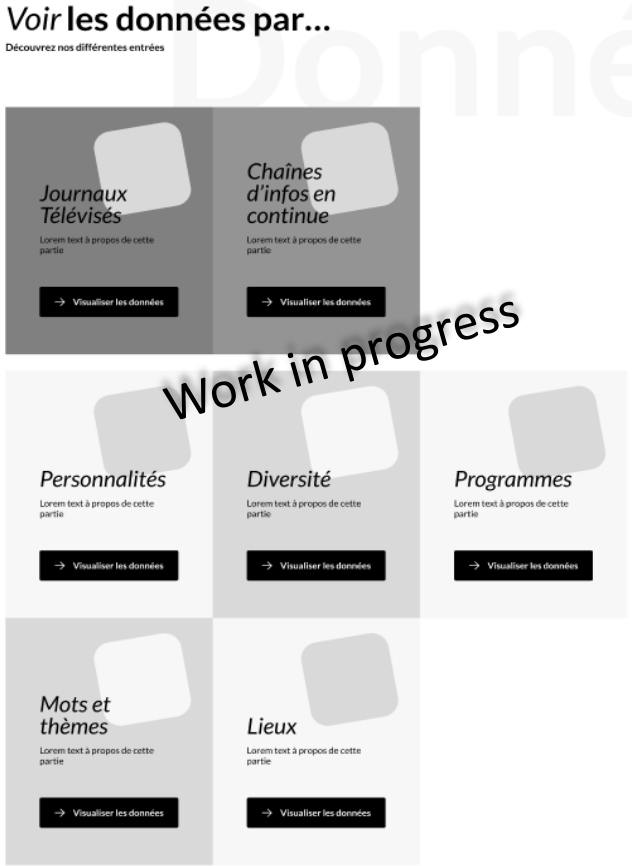
Des tableaux de bord interactifs pour proposer des chiffres dans la durée sur les mêmes indicateurs

Exemple de tableau de bord sur les personnalités



Exemple de tableau de bord sur les personnalités

Les données d'indicateurs sont téléchargeables et disposent d'une licence ouverte pour en permettre la réutilisation

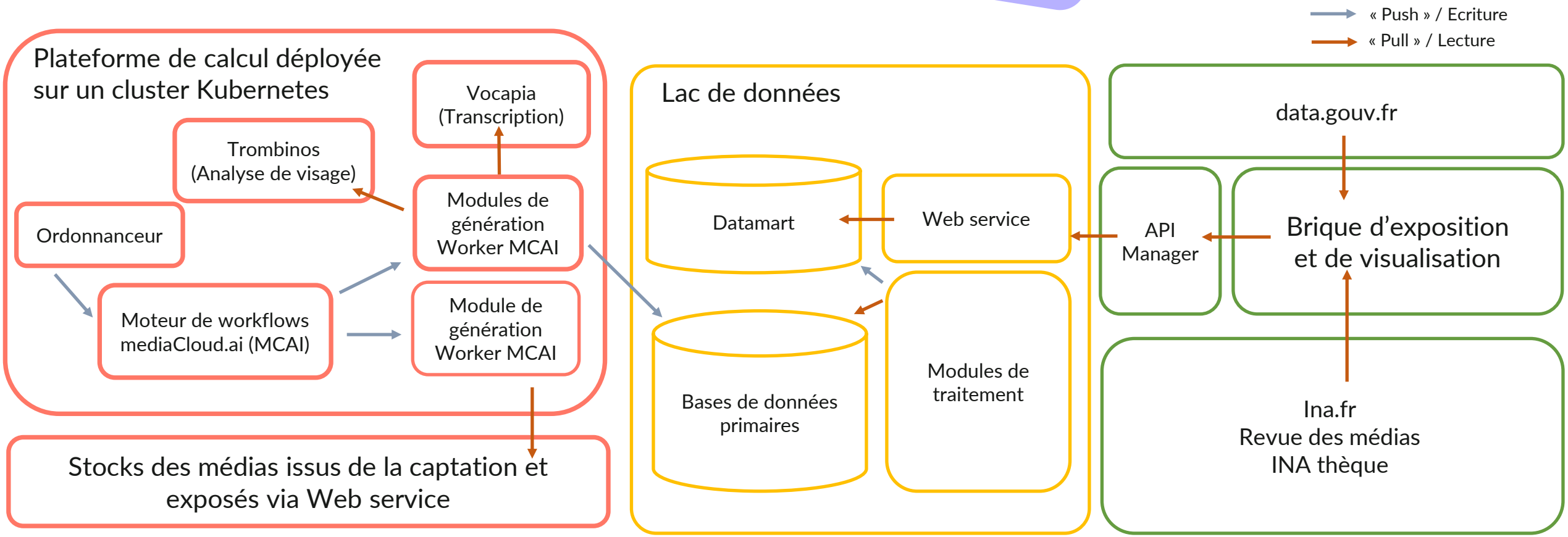


Les indicateurs sont répartis selon deux axes :

- des corpus
  (journaux télévisés, chaînes d'information continue et programmes pour la V1)
- des thématiques
  (Personnalités, diversité, lieux, mots et thèmes pour la V1)

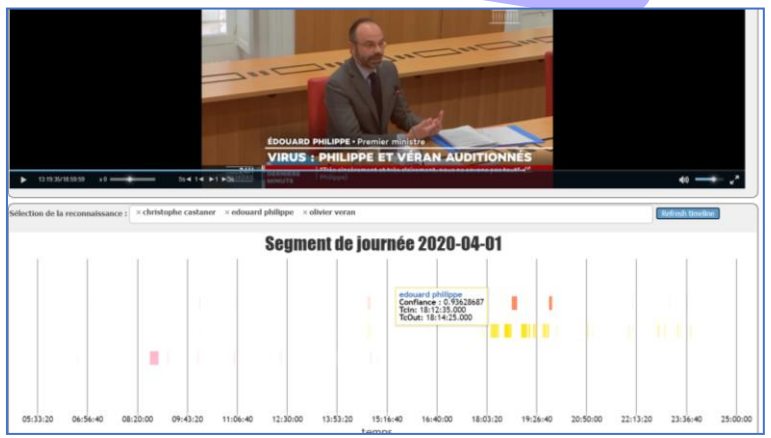
Le travail de design, de maquettage et d'UX/UI est effectué par la société WeDoData

Architecture de la solution

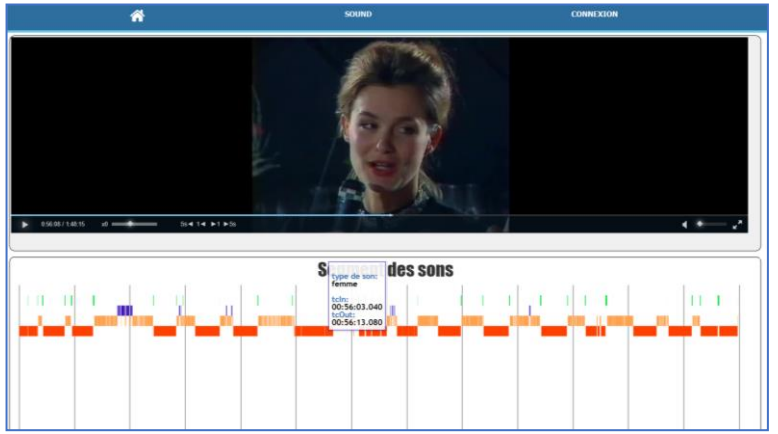


Création des données (les algos)

Transcription effectuée avec Vocapia et extraction d'entités nommées avec TextRazor et reliés à notre référentiel via un alignement avec Wikidata



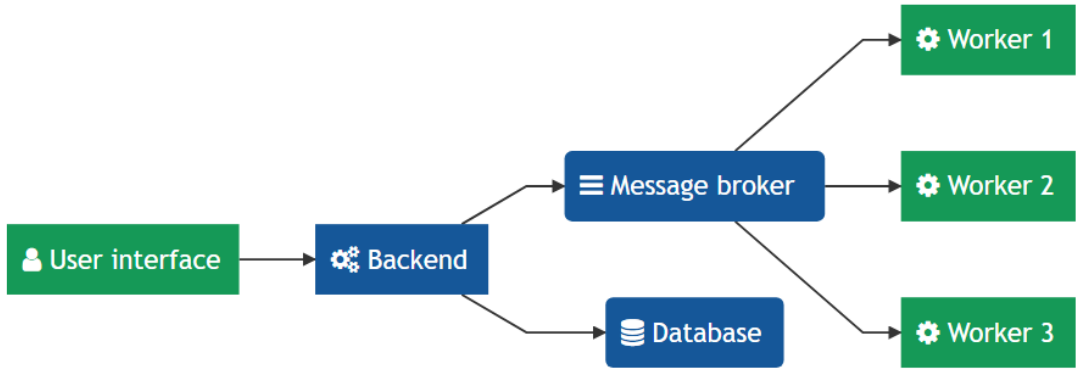
Identification automatique des visages de + 70 000 personnes avec ***Trombinos*** mis au point par
Pierre Letessier, chercheur à l'INA



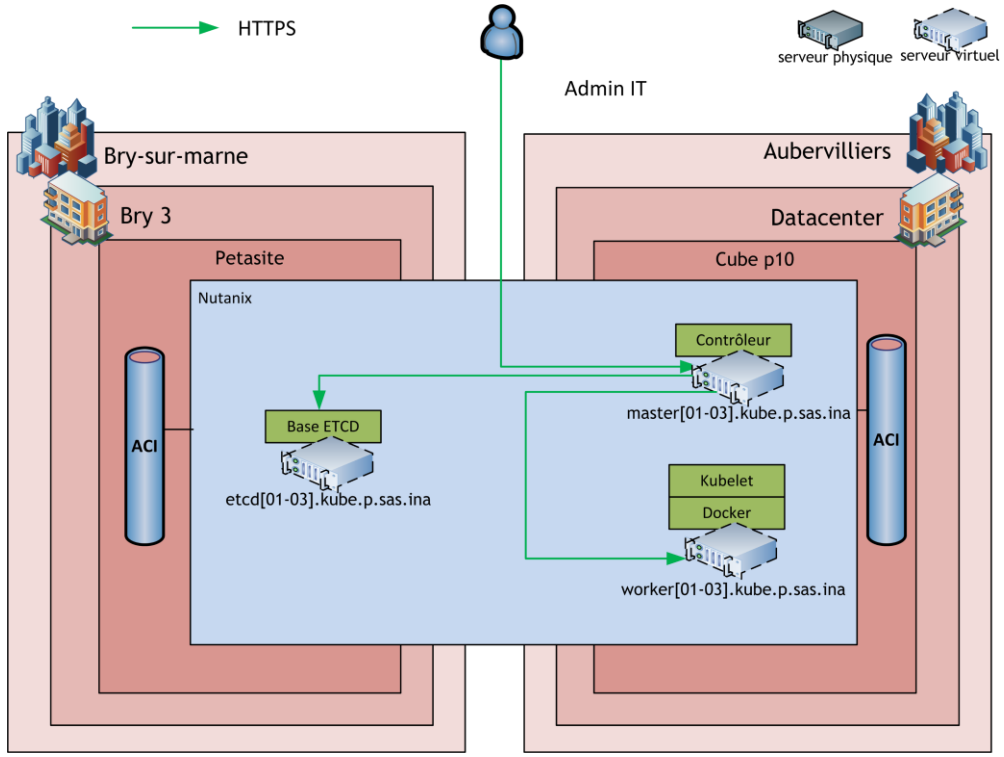
Analyse du son (homme, femme, bruit, musique) avec InaSpeechSegmenter mis au point par David Doukhan, chercheur à l'Ina

On peut aussi citer : l'OCR effectué avec la bibliothèque Open Source EasyOCR, classification automatique des images avec la solution Deepomatic, traitement automatique du langage avec les frameworks Transformers et Tensor Flow, la segmentation automatique de programme, la détection de jingles, le repérage des visages et la reconnaissance d'images avec OpenCV et bien-sûr l'ensemble des métadonnées et référentiels mis au point depuis 46 ans par les professionnels de l'information de l'Ina…

Création des données (la plateforme)

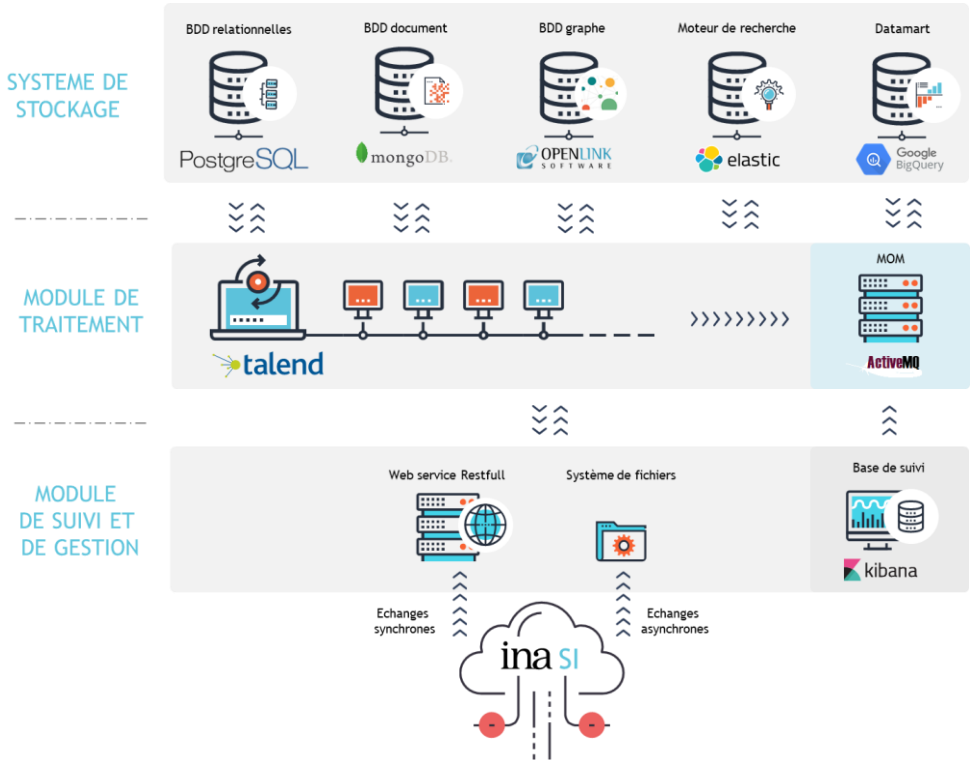


L'ensemble des traitements sont orchestrés par le moteur de workflow, Media Cloud AI, plateforme Open Source de micro-services scalable et distribuable, mis au point et maintenu par France Télévision et la société Media-IO.

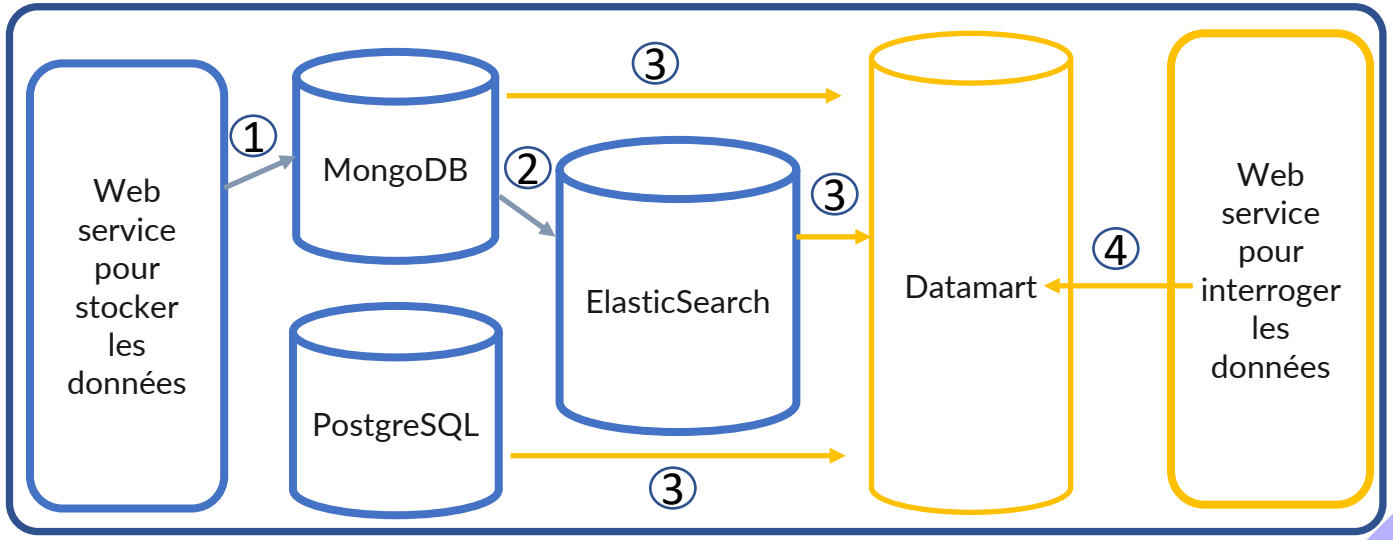


La plateforme est installée sur un cluster Kubernetes déployé sur nos propres infrastructures réparties sur deux sites physiques.

Stockage et traitement des données

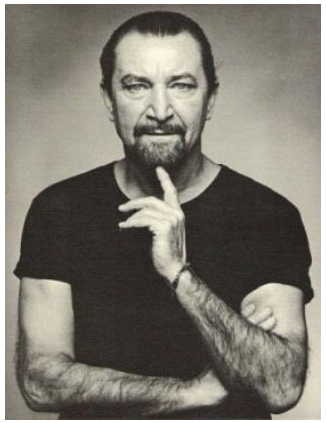


Une infrastructure pour stocker, mettre en cohérence, traiter et exposer toutes les données de l'INA



Les différents rôles

Architecte
de données



Data scientist

Ingénieur
de la donnée

Analyste de la donnée
Professionnels de l'information

Data ops/Dev Ops